CA4023

Assignment One

Semester Two, 2023

PART ONE (4.5 marks)

Sentence Boundary Detection

Write a program that splits a document into sentences. The input to your program should be a file containing text. The output should be a new file with each sentence from the first file on a separate line.

For example, if the input file contains the following:

The difference for them can be significant. A 10-year loan of €50,000 from you at zero interest would involve monthly payments of €417, or €5,000 a year. With An Post's market-leading rate, the friend would be paying €525.45 a month, or over €6,300 a year and over €13,000 interest over the life of the loan.[1]

The output file should contain the following:

The difference for them can be significant.
A 10-year loan of €50,000 from you at zero interest would involve monthly payments of €417, or €5,000 a year.
With An Post's market-leading rate, the friend would be paying €525.45 a month, or over €6,300 a year and over €13,000 interest over the life of the loan.

Note that your solution should NOT make use of machine learning.

You may use a library for regular expressions.

[1] Excerpt from https://www.irishtimes.com/your-money/2023/01/29/lending-to-family-and-friends-can-leave-you-very-exposed/

PART TWO (4.5 marks)

Language Modelling

Implement an unsmoothed bigram language model. Train your model on the following toy corpus:

```
<s> b c </s>
<s> c c </s>
<s> c b </s>
<s> b b </s>
```

Calculate and print out the probability of each of the following strings:

```
<s> c </s>
<s> b </s>
<s> b c </s>
<s> b b </s>
<s> b c b </s>
```

PART THREE (10 MARKS)

Naive Bayes Sentiment Polarity Classifier

Write a sentiment polarity classifier which uses the Naive Bayes algorithm to train a sentiment polarity classifier which assigns a sentiment polarity of `positive` or `negative` to a review.

Your program should accept as input a training file and a test file. The training file contains a list of reviews and their actual sentiment labels (`positive` or `negative).` The test file contains either a list of reviews with the actual sentiment labels or a list of the reviews on their own. Your program should output the predictions of the NB classifier (`positive` or `negative)` for each of the reviews in the test file. If the actual labels (sometimes referred to as gold labels or ground truth) are also available for the test reviews, your program should also print the accuracy of the classifier.

You should use the following training data:

https://www.cs.cornell.edu/people/pabo/movie-review-data/review_polarity.tar.gz

described in the following paper:

Pang and Lee 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *Proceedings of the 42nd ACL.*
https://www.aclweb.org/anthology/P04-1035/

There are 1000 positive reviews and 1000 negative reviews. Reserve the last 100 of each type for testing (files starting with *CV9*) and the first 900 for training (files starting with *CV[0-8]*).

Your program should be case-insensitive and should ignore punctuation.

Analyse the output of your classifier on 5 correct and 5 incorrect samples from the test set. For each example, say why you think your classifier made the correct or incorrect decision.

Points to Note

- You may implement the solutions in a programming language of your choice.
- Note that you may NOT make use of external NLP or machine learning libraries.

Marking Criteria

For Parts One and Two, marks will be awarded for

1. Correct implementation (4 marks)
2. Clear, readable, appropriately commented code (0.5 marks)

For Part Three, marks will be awarded for

1. Correct implementation (7 marks)
2. Clear, readable, appropriately commented code (1 mark)
3. Analysis (2 marks)

The remaining mark will be awarded for correctly forking the master git repo (see below) and uploading your assignment files to the correct location of your forked repo.

Assignment Submission

1. School of Computing Gitlab repository. Follow the README instructions for forking the master repository, 2023-ca4023-master
2. Each folder (part-one, part-two and part-three) of your forked repo should contain the code and a README.md describing how to execute the program. The folder for Part Three should also contain a file called ANALYSIS.md which contains the analysis of the 10 randomly chosen samples.

Assignment Deadline

Fri 24th Feb, 6pm